模式识别-第二次作业

王健
201628015029018

1、第一题

1) 对于一维特征空间中的 Parzen 窗估计，其估计得到的概率密度函数为：

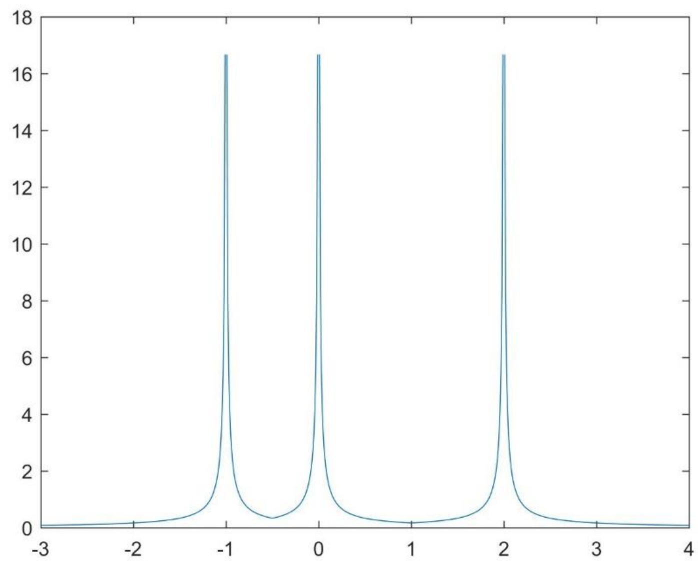
$$p_n(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi(\frac{\mathbf{x}-\mathbf{x}_i}{h_n})$$

2) 对于一维特征空间中的最近邻估计，得到的概率密度函数为：

$$p_n(\mathbf{x})=\frac{1}{2nh_{\mathbf{x}}}$$

其中 $h_{\mathbf{x}}$ 表示 x 点与最近训练样本数据点之间的距离。

根据所给出的数据点，概率密度函数做图如下：



2、

（a）E step：

$$
\begin{aligned}
Q(\boldsymbol{\theta};\boldsymbol{\theta}^0) &= \mathrm{E}_{x_{32}}[\ln p(\mathbf{x}_g,\mathbf{x}_b;\boldsymbol{\theta}) \mid \boldsymbol{\theta}^0, D_g] \\
&= \int_{-\infty}^{\infty} (\ln p(\mathbf{x}_1 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_2 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_3 \mid \boldsymbol{\theta}))\, p(x_{32} \mid \boldsymbol{\theta}^0, x_{31}=2)\, dx_{32} \\
&= \ln p(\mathbf{x}_1 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_2 \mid \boldsymbol{\theta}) + \int_{-\infty}^{\infty} \ln p(\mathbf{x}_3 \mid \boldsymbol{\theta})\, p(x_{32} \mid \boldsymbol{\theta}^0, x_{31}=2)\, dx_{32} \\
&= \ln p(\mathbf{x}_1 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_2 \mid \boldsymbol{\theta}) + \int_{-\infty}^{\infty} \ln p(\begin{pmatrix} 2 \\ x_{32} \end{pmatrix} \mid \boldsymbol{\theta}) \frac{p(\begin{pmatrix} 2 \\ x_{32} \end{pmatrix} \mid \boldsymbol{\theta}^0)}{\int_{-\infty}^{\infty} p(\begin{pmatrix} 2 \\ x'_{32} \end{pmatrix} \mid \boldsymbol{\theta}^0)\, dx'_{32}}\, dx_{32} \\
&= \ln p(\mathbf{x}_1 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_2 \mid \boldsymbol{\theta}) + \int_{-\infty}^{\infty} \ln p(\begin{pmatrix} 2 \\ x_{32} \end{pmatrix} \mid \boldsymbol{\theta})\, p(\begin{pmatrix} 2 \\ x_{32} \end{pmatrix} \mid \boldsymbol{\theta}^0)\, dx_{32} \\
&= \ln p(\mathbf{x}_1 \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_2 \mid \boldsymbol{\theta}) + K
\end{aligned}
$$

K有三种情况：

$1.3 \le \theta_2 \le 4$:

$$K = \frac{1}{4}\int_0^{\theta_2} \ln(\frac{1}{\theta_1} e^{-2\theta_1} \frac{1}{\theta_2})\, dx_{32} = \frac{1}{4}\theta_2 \ln(\frac{1}{\theta_1} e^{-2\theta_1} \frac{1}{\theta_2})$$

$2.\theta_2 \ge 4$:

$$K = \frac{1}{4}\int_0^{4} \ln(\frac{1}{\theta_1} e^{-2\theta_1} \frac{1}{\theta_2})\, dx_{32} = \ln(\frac{1}{\theta_1} e^{-2\theta_1} \frac{1}{\theta_2})$$

$3.otherwise$

$K = 0$

(b)最大化 Q：

$1.3 \le \theta_2 \le 4$:

$$Q(\theta;\theta^0) = -4 - (2\ln\theta_2 + \frac{1}{4}\theta_2(2+\ln\theta_2))$$

这个函数对于θ来说是单调的，则当 Q 最大的时候，$\theta_2=3$，此时 Q=-8.52

$2.\theta_2 \ge 4$:

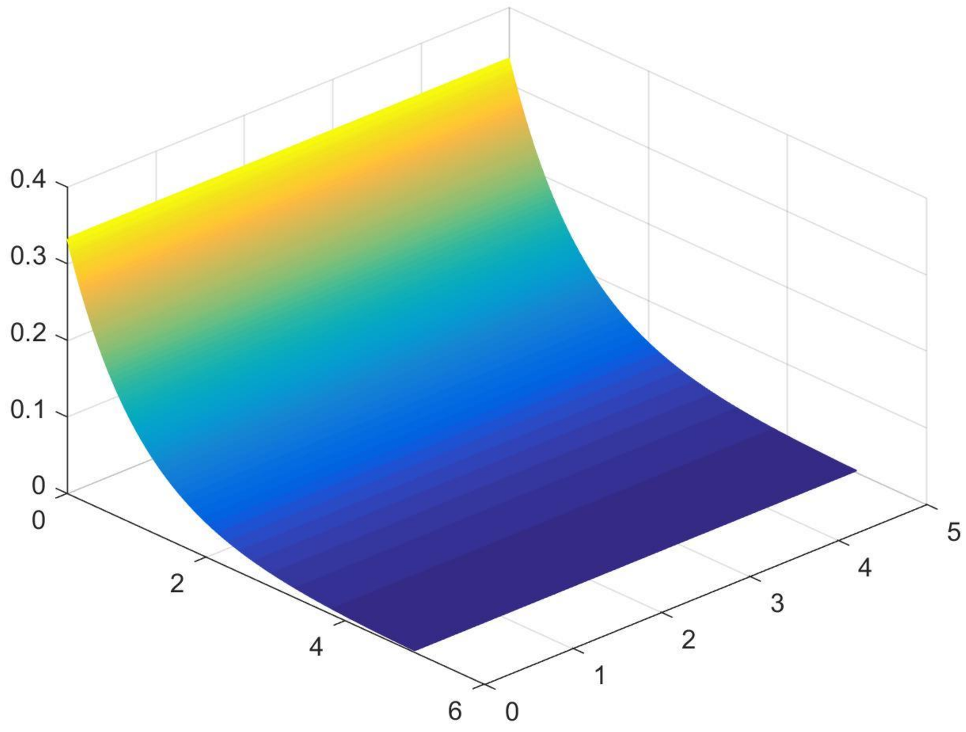
$$Q(\theta;\theta^0) = -6 - 3\ln\theta_2$$

当 Q 最大的时候，$\theta_2=4$，此时 Q=-10.16

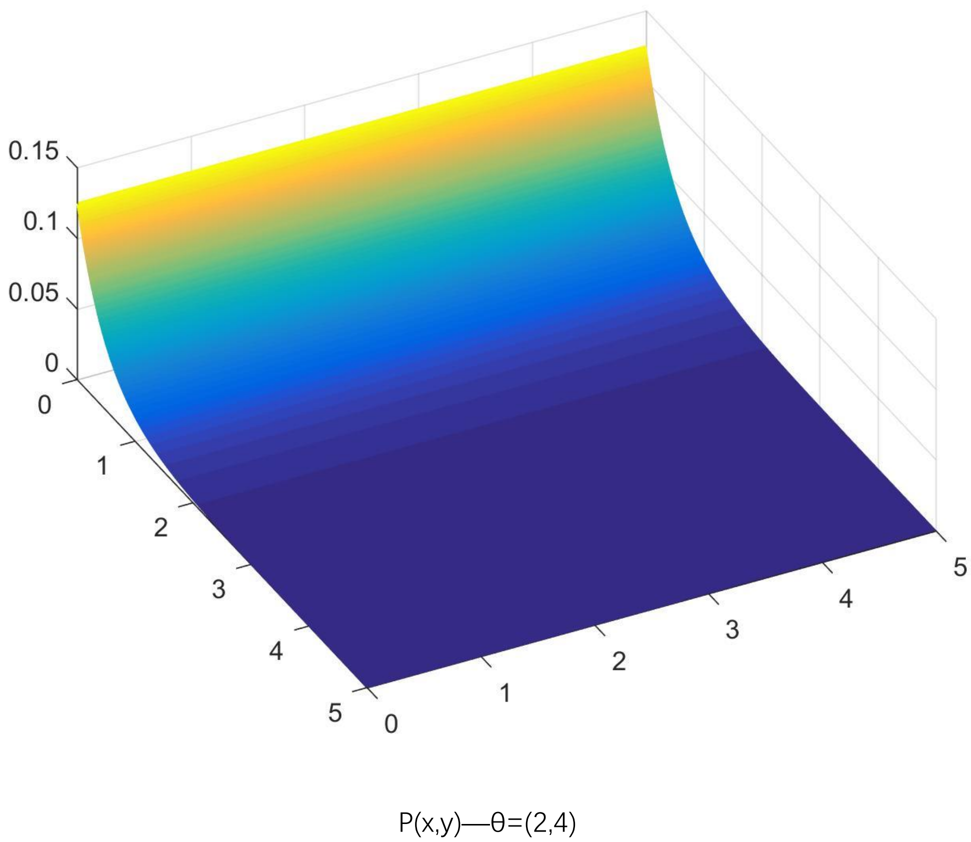
$\theta_{32} = 3$

综合来说，$\boldsymbol{\theta} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

（c）两个函数的图像为：



P(x,y)—θ=(1,3)



P(x,y)—θ=(2,4)

3、(1) 表达式为：

$$P(z_{i1},...,z_{in} \mid P(\omega_i)) = \prod_{k=1}^{n} p(\mathbf{z}_{ik} \mid P(\omega_i))$$

(2) 最大似然估计：

$$\max_{\theta} P(z_{i1},...,z_{in} \mid P(\omega_i)) \leftrightarrow \frac{d(\prod_{k=1}^{n} p(z_{ik} \mid P(\omega_i)))}{d(P(\omega_i))} = 0$$

$$P(\omega_i) = \frac{\sum_{k=1}^{n} z_{ik}}{n}$$

4、(1) 1-NN 规则：

对于测试样本点 x，在已知的训练集合中寻找距离它最近的点，记为 x‘，那么将点 x 分为 x‘所属的类别

3-NN 规则：对于测试样本点，在已知的训练集合中距离最近的三个点，记为 x_1,x_2,x_3，三个点中若存在两个或两个以上点属于某一类别，那么判别 x 为那个类别

(2) 重新设计：

1-NN 规则：

对于测试样本点 x，在已知的训练集合中寻找距离它最近的点，记为 x‘，那么将点 x 分为 x‘所属的类别，如果有两个或者多个不同类别的点与 x 距离相同且最近，那么拒绝判别

3-NN 规则：对于测试样本点，在已知的训练集合中距离最近的三个点，记为 x_1,x_2,x_3，三个点中若存在两个或两个以上点属于某一类别，那么判别 x 为那个类别。如果在距离点 x 前三近的点不止一个，那么可以将它们都考虑进来，设为 x_1,x_2,…,x_n，如果有恰好半数个点属于某一类，另外半数个点属于另一类，那么拒绝判别，否则则判别为多数的类别。

(3) 优缺点：

1、优点

简单，易于理解，易于实现，无需估计参数，无需训练

适合对稀有事件进行分类

特别适合于多分类问题

2、缺点

对测试样本分类时的计算量大，内存开销大

可解释性较差

5、本次使用了 1-NN，3-NN，5-NN，三种方法对图像进行了分类，代码为 python，程序见附件

三种分类方法的分类结果如下表所示：

表 1　分类方法及其准确率

分类方法	准确率
1-NN	0.9691
2-NN	0.9713
3-NN	0.9694

可以看出即使是最近邻方法，其分类准确率也相当高，而 3-NN 和 5-NN 对于准确率的提升效果不明显，而 5-NN 判断的准确率反而有所下降，可能是由于测试数据集中某些经过旋转或者平移的数字，与其欧氏距离近的数字类别多，例如 5 个最近的样本点的类别为[1,2,3,4,5]，在类似此种情况下，在欧式距离最近的 5 个样本点中选取最多的那个类别，则不是一个很好的策略，在这种情况下，选择距离最近的样本点所属的类别，可能相对比较准确。
针对 1-NN 方法进行分析，对于不同的数字分类准确率为：

表 2 不同数字的分类准确率

数字	准确率
0	0.9929
1	0.9947
2	0.9612
3	0.9604
4	0.9613
5	0.9641
6	0.9854
7	0.9650
8	0.9446
9	0.9584

从表格中可以看出，数据对 0,1 此类容易分辨的数字，分类效果相当良好，但是对于 8，9 这种本身难于分辨的数字，分辨效果就稍差，考虑到在测试数据集中存在数字图像位置不在中心，数字图像有少许旋转等情况，分辨效果稍差也是可以理解的。